International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

Early Detection of Parkinson's Disease Using Voice and Speech Analysis

Vivek Ugale[1], Prof. Gaurav Malode[2], Sanket Rindhe[3], Abhijeet Shinde[4], Vedant Mali[5]
[1,2,3,4,5]*Dept. of Information Technology, International Institute of Information Technology, Pune, India 411057*

***Abstract*— Parkinson's Disease (PD) is a progressive neurological condition that disrupts motor control and speech production, with vocal dysfunction (dysphonia) affecting up to 89% of patients often preceding motor symptoms by several years. This paper presents a comprehensive voice-based PD detection system combining a 149-dimensional acoustic and clinical feature pipeline with multiple classification frameworks. The feature vector integrates Mel-Frequency Cepstral Coefficients (MFCC), temporal derivatives, spectral features, and clinically validated Praat biomarkers Jitter, Shimmer, HNR, and F0 statistics which are directly associated with vocal fold dysfunction in PD. Two centralized models trained and evaluated using rigorous 5-fold person-wise cross-validationSupport Vector Machine (SVM) with RBF kernel and Random Forest achieved segment-level accuracies of 98.40% (±2.47%) and 97.60% (±2.04%) respectively, with AUC-ROC above 0.996. A lightweight Convolutional Neural Network (LightCNN) operating on mel spectrograms provides an efficient deep learning alternative. A Federated Learning Neural Network trained across three private datasets using weighted FedAvg achieved 95.53% accuracy with a clinically critical sensitivity of 98.53%, without centralizing any patient data. SHAP analysis provides feature-level clinical explainability, confirming that Shimmer, HNR/NHR, and MFCC coefficients are the dominant discriminators. A systematic ablation study validates the contribution of each feature group, with clinical Praat features providing measurable gain over acoustic-only baselines.**

***Keywords*—Acoustic Features, CNN, Federated Learning, FedAvg, MFCC, Parkinson's Disease, Praat Biomarkers, SHAP Explainability, Support Vector Machine, Voice Analysis.**

I. Introduction

Parkinson's Disease (PD) ranks as the world's second most common neurodegenerative disorder, affecting more than 11.67 million people globally as of 2023 a figure that has risen by 37% since 1990 [16]. It is characterized by progressive loss of dopaminergic neurons in the substantia nigra, leading to motor symptoms including tremor, rigidity, and bradykinesia, as well as non-motor effects such as cognitive impairment, depression, and speech dysfunction. PD is most prevalent in people over 60, and with an ageing global population, incidence is projected to rise substantially in coming decades.

One of the most widespread yet underappreciated manifestations of PD is vocal dysfunction. Studies consistently show that dysphonia encompassing reduced vocal intensity, flat monotone pitch, breathiness, and voice tremor affects up to 89% of PD patients [10, 13]. Critically, vocal changes can appear years before motor symptoms become clinically apparent, making acoustic analysis a promising and non-invasive avenue for early-stage screening. Yet only 3–4% of PD patients receive speech therapy [13], reflecting the reliance on subjective clinical assessment and the shortage of accessible specialist services.

Standard PD diagnosis relies on neurological evaluation of motor symptoms, which typically do not manifest until 60–80% of dopaminergic neurons have already been lost [15]. At early stages, clinical accuracy sits at only 40–50% [15]. Voice changes measurable through acoustic analysis offer a fundamentally different and earlier detection window. Automated voice-based screening systems are scalable, non-invasive, and affordable requiring only a basic microphone recording.

Deploying such systems across multiple healthcare institutions introduces a critical challenge: patient privacy.

Voice recordings carry sensitive health information, and cross-institutional data sharing is tightly regulated under HIPAA, GDPR, and equivalent frameworks. Federated Learning (FL) addresses this directly models train across distributed private data sources by sharing only weight updates, never raw recordings [9].

II. Background And Literature Review

A. Pathophysiology of Voice Impairment in Parkinson's Disease

Human voice is shaped by three tightly linked physiological systems: respiration, laryngeal phonation, and articulatory coordination. In PD, dysfunction in the basal ganglia disrupts the fine neural coordination that all three systems rely upon, resulting in hypokinetic dysarthria a speech pattern marked by reduced intensity, flat monotone pitch, imprecise articulation, and irregular voice tremors [14].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

At the acoustic signal level, these changes produce measurable biomarkers: Jitter (cycle-to-cycle fundamental frequency variation) rises due to irregular vocal fold movement; Shimmer (amplitude variation between cycles) increases as laryngeal control weakens; and Harmonic-to-Noise Ratio (HNR) falls, capturing growing breathiness. Together, these perturbation measures form the clinical basis for voice-based PD detection [8].

B. Literature Survey

The foundation for voice-based PD detection was established by Little et al. [1], who demonstrated that dysphonia measures fed to an SVM classifier could reliably separate PD from healthy voice, reaching approximately 91% accuracy on a 22-person dataset using a 31-dimensional feature set. Sakar et al. subsequently expanded this approach with MFCC features, achieving 96.2% on a controlled vowel dataset. Alshammri et al. [6] demonstrated that combining acoustic perturbation measures with MFCCs in an SVM achieves 98% accuracy and 99% F1-score. Di Cesare et al. [7] applied MFCC and GTCC features to spontaneous dialogue and achieved 92.3% with speaker diarization an important step toward naturalistic speech settings. On the deep learning side, CNN architectures applied to mel spectrograms have demonstrated competitive performance by capturing temporal and spectral patterns that fixed feature vectors may not fully represent.

For federated learning, Sarlas et al. [5] adapted leading centralized PD detection to a federated framework, demonstrating competitive accuracy without raw data sharing. Leitner et al. [2] trained a cross-lingual FL model on German, Spanish, and Czech speech data, finding that federated aggregation can actually improve generalization beyond individually trained local models. Table 1 presents a comparative summary of related studies.

TABLE I
COMPARATIVE SUMMARY OF RELATED STUDIES ON VOICE-BASED PARKINSON'S DISEASE DETECTION

Study	Method	Dataset / Features	Accuracy
Little et al. [1]	SVM	Oxford dataset, dysphonia features	~91%
Sarlas et al. [5]	Federated DL	Multi-institutional speech data	FL competitive with centralized
Di Cesare et al. [7]	SVM + MFCC + GTCC	Spontaneous dialogue task	92.3%
Alshammri et al. [6]	SVM + MFCC	Combined acoustic features	98%
Proposed System	SVM / RF / FL	3 heterogeneous datasets, 149 features, Praat biomarkers	98.40% SVM, 98.00% FL-NN

Note: FL = Federated Learning; SVM = Support Vector Machine; RF = Random Forest; GTCC = Gammatone Cepstral Coefficients

C. Research Gaps and Motivation

Despite this progress, key limitations persist in existing work. Most studies evaluate on a single dataset, restricting generalizability. Subject-wise cross-validation critical for preventing data leakage in segmented audio is not consistently applied, causing many reported accuracies to be optimistic. Clinical explainability connecting model decisions to known vocal biomarkers is rarely provided. And federated approaches rarely use genuinely heterogeneous multi-source data with different languages, recording conditions, and demographics.

This work directly addresses these gaps: training and evaluating across three structurally distinct datasets simultaneously, enforcing strict person-wise validation, incorporating Praat clinical biomarkers alongside standard acoustic features, providing SHAP explainability, and implementing Weighted FedAvg to handle large differences in client dataset size.

III. SYSTEM REQUIREMENTS AND DATASET ANALYSIS

A. Dataset Description

Three publicly available datasets were used to simulate a realistic multi-institutional federated learning scenario with heterogeneous Non-IID distributions. They differ fundamentally in recording environment, language, phonation task, and demographics making the combination a genuinely challenging test for both centralized and federated models. Table II summarizes their key characteristics.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

TABLE II
. DATASET SUMMARY — THREE HETEROGENEOUS PD SPEECH DATASETS

Dataset	Patients (PD / HC)	Recording Type	Seg. Duration	Orig. Segments	Post-Aug. Segments
Dataset 1 Italian	24 PD / 22 HC	Speech, vowels, breathing	3 sec	5,152	20,608
Dataset 2 Controlled vowel /a/	40 PD / 41 HC	Sustained vowel /a/	1 sec	228	912
Dataset 3 Large multi-source	333 PD / 277 HC	Continuous speech	3 sec	654	2,616
Combined Total	397 PD / 340 HC (727 total)	Heterogeneous (Non-IID)	—	6,034	24,136

B. Software and hardware requiremnets

The proposed system was developed using Python 3.8 as the main programming language. Key software libraries include Librosa for loading audio, normalization, and extracting acoustic features like MFCC and spectral features. Parselmouth, a Python wrapper around Praat, is used for extracting clinical biomarkers such as Jitter, Shimmer, HNR, and F0. Scikit-learn is utilized for SVM, Random Forest, StandardScaler, and cross-validation. PyTorch supports the Federated Neural Network and LightCNN architectures. SHAP is used for explaining model decisions, while Streamlit provides the interactive web deployment interface. Missing Praat feature values, which account for 0.9% of segments, were processed using Scikit-learn's SimpleImputer with the median strategy. All source code was developed in Visual Studio Code.

The entire system, which includes building the dataset, extracting features, training the model, performing SHAP analysis, and conducting federated learning, was run on a standard consumer-grade CPU machine without needing any GPU acceleration. The minimum hardware requirement is an Intel Core i5 processor from the 8th generation or equivalent, along with 8 GB of RAM and 50 GB of available storage. A GPU is optional and only helpful for training CNN-based spectrograms. The three voice datasets together take up about 5 GB of storage. The complete end-to-end training process, from raw audio to trained models, takes around 40 to 60 minutes on CPU.

IV. SYSTEM DESIGN AND ARCHITECTURE

A. Overall System Pipeline

The proposed system uses a step-by-step process to turn voice recordings into a clinical prediction. This process is the same for all model types, which helps ensure that the data is prepared and features are extracted consistently. The process starts with audio files in WAV format recorded at 16 kHz. Each recording is first adjusted to remove differences in loudness. This is done using a tool called librosa. Then silent parts are removed from the recordings. This helps get rid of noise that could interfere with the analysis.The cleaned audio is then broken down into parts. For some datasets these parts are 3 seconds long. For another they are 1 second long. This helps create data for training while keeping the important parts of the recordings intact.Each part is then analyzed to extract features. This produces a list of 149 numbers that describe the audio. These features include things like MFCC and Praat biomarkers, which are used in clinics. If there are any values they are replaced with average values from the training data. The features are then adjusted to have scales.

The prepared features are then used to make a prediction. When a recording is analyzed each part is looked at separately. The predictions for each part are then combined to make a prediction, for the whole recording. The output includes a predicted class, which's either PD or Healthy. It also includes a confidence score and a risk level, which can be Low, Medium or High.The complete process is shown in Fig. 1.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

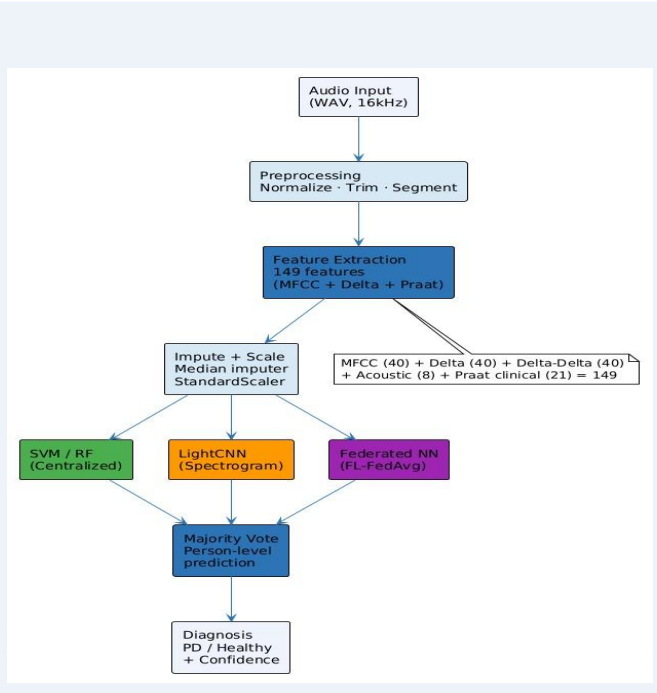

Fig. 1. System Architecture Pipeline-- Complete Flow from Audio Input to Final PD Prediction. The pipeline is shared across all model types; the preprocessing and feature extraction stages are identical for centralized and federated models.

B. Centralized Architecture (SVM and Random Forest)

The centralized architecture is used to train and evaluate models on a combined dataset from all three sources on a machine. We use two machine learning models as the primary classifiers: Support Vector Machine and Random Forest.

Support Vector Machine is used because it is good at handling dimensional feature spaces, which is what we have with our 149-dimensional feature vector. It also gives us calibrated probability estimates, which we need for majority voting at the person level. The way it works is that it maps the input features into a dimensional space where it can separate the two classes even if the decision boundary is not linear in the original feature space.

We selected the Support Vector Machine with an RBF kernel as the deployment model. The RBF kernel is used with the following settings: C equals 1.0 gamma equals 'scale'. Probability equals True.

Random Forest is another model we use. It uses an ensemble of 200 decision trees that are trained with the Gini impurity criterion. Each tree is trained on a subset of the training data and a random subset of features at each split. This helps reduce variance and improve generalization. Random Forest is also useful because it gives us feature importance scores, which we use in the ablation study and it is compatible with SHAP TreeExplainer for Shapley value computation. This allows us to do per-prediction explainability analysis.

We evaluate both models using 5fold person-wise cross-validation. This means we split the data into five folds based on person IDs so that no speaker appears in both the training and test sets. This is important because it prevents speaker identity leakage, which can inflate performance. Within each fold we fit the median imputer. Standardscaler exclusively on the training split and apply them to the test split so that no information from the test set influences preprocessing. We get person-level predictions by taking a majority vote across all segment predictions for each patient. We use the mean probability for AUC-ROC computation. Finally we train the models on the full dataset and save them for deployment, in the Streamlit web application.

C. Federated Learning Architecture

The federated setup simulates a privacy-preserving multi-institutional environment with three independent clients and one central server. The global model is a 4-block neural network (149→256→128→ 64→2) with BatchNorm1d, Dropout regularization, and a residual skip connection. BatchNorm stabilizes activations across Non-IID client distributions; the residual connection prevents gradient degradation over 10 training rounds.

Each training round follows: (1) server distributes current global weights to all clients; (2) each client trains locally for 5 epochs using Adam optimizer on its private data; (3) clients return updated weights; (4) server aggregates using Weighted FedAvgGlobal Weights = $\Sigma(W_i \times n_i) / \Sigma n_i$ weighting each client proportionally to its dataset size. Each client maintained independent median imputer and StandardScaler fitted on local training data only, preventing cross-client leakage

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

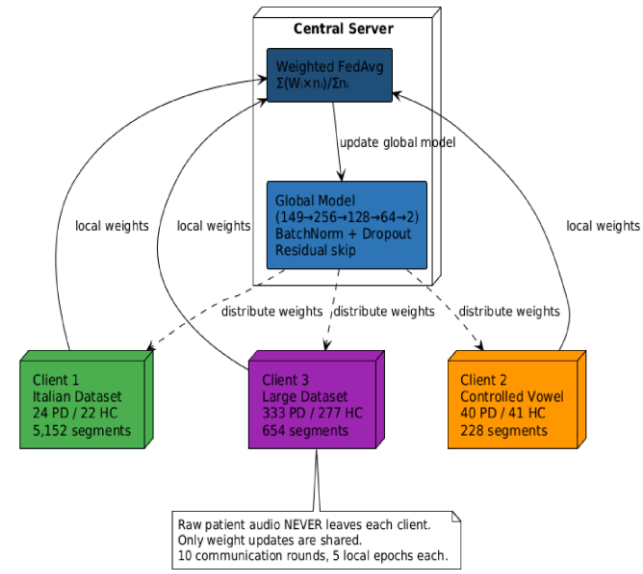

Fig. 2. Federated Learning Architecture — Three-Client Simulation with Weighted FedAvg Aggregation. Dashed arrows indicate global weight distribution; solid arrows indicate local weight upload. Raw patient audio never leaves each client.

V. IMPLEMENTATION

A. Audio Preprocessing and Segmentation

Raw .wav files are loaded at 16 kHz using Librosa. Amplitude normalization is applied to equalize loudness differences across recordings. Silence is removed using a 20 dB top-db threshold (silence relative to peak energy). Each file is then segmented into fixed-length windows 3 seconds for Datasets 1 and 3, 1 second for Dataset 2. Segments shorter than the target length are zero-padded. At inference time, each segment is classified independently, and segment predictions are combined by majority probability voting to produce a single person-level diagnosis.

Critically, augmentation was applied only in the extended pipeline: Gaussian noise addition (SNR ≈30 dB), time stretching (rate 0.9), and pitch shifting (+2 semitones).

All augmented segments retain the original speaker's person_id, ensuring they remain in the same cross-validation fold as their source and preventing augmentation-based data leakage. Clinical features (Jitter, Shimmer, HNR) for augmented segments were reused from the original to avoid unreliable Praat extraction on modified audio. Fig. 3 illustrates the augmentation pipeline design.

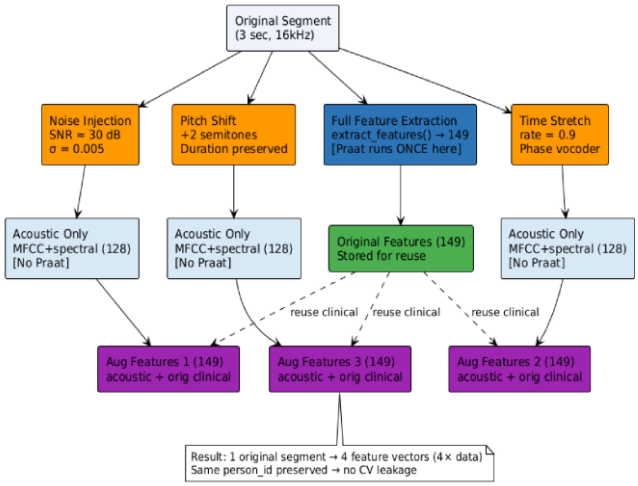

Fig. 3. Data Augmentation Pipeline. Praat (clinical features) runs once per original segment only. Augmented variants reuse original clinical features and preserve speaker identity — yielding 4× data volume without leakage.

B. Feature Extraction — 149 Features

For every audio segment, a 149-dimensional feature vector is computed. All acoustic features are extracted with Librosa using hop_length=512, n_fft=2048. Clinical biomarkers (Jitter, Shimmer, HNR, NHR, F0 statistics) are extracted using parselmouth a Python wrapper around Praat, the clinical-grade voice analysis tool validated in PD research since Little et al. [1]. Of 5,152 segments, 91 NaN values occurred in features 132–140 (Jitter/Shimmer, 1.77% of segments), corresponding to segments too short or noisy for reliable Praat extraction. These were handled by median imputation fitted on training data only consistent with standard practice and preventing test set leakage. Table 4 details the full feature composition

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

TABLE III
149-DIMENSIONAL FEATURE VECTOR COMPOSITION

Feature Group	Count	Description / Clinical Relevance
MFCC mean & std (20 coeff × 2)	40	Spectral envelope of vocal tract
Delta MFCC mean & std	40	First-order temporal derivatives of MFCC
Delta-Delta MFCC mean & std	40	Second-order derivatives (spectral acceleration)
Pitch, ZCR, Spectral Centroid, RMS (mean+std)	8	Fundamental acoustic properties per segment
Jitter: local, RAP, PPQ5, DDP	4	Cycle-to-cycle F0 variation laryngeal tremor in PD
Shimmer: local, APQ3, APQ5, APQ11, DDA	5	Amplitude variation vocal fold instability in PD
HNR + NHR (Praat)	2	Harmonics-to-Noise ratio voice breathiness hallmark
F0 statistics: mean, std, min, max	4	Pitch range and stability reduced in PD
Spectral Flux, Rolloff, Bandwidth (mean+std)	6	Frequency distribution irregularity over time
Total	149	Concatenated fixed-length feature vector per segment

C. Machine Learning Models

Support Vector Machine (SVM).

SVM uses Scikit-learn's SVC with an RBF kernel (C=1.0, gamma='scale') and probability=True for Platt scaling. SVM is selected as the primary centralized model for deployment due to its strong performance on high-dimensional feature spaces and support for probability calibration enabling majority voting at the person level.

Random Forest (RF).

Random Forest uses 200 estimators with Gini impurity criterion.

In addition to classification, Random Forest produces feature importance scores used in the SHAP and ablation analyses. Its ensemble structure provides natural robustness to the mild class imbalance (62.1% PD / 37.9% HC) in the dataset.

D. Federated Learning Implementation

The federated model is a 4-block MLP (149→256→128→64→2) with BatchNorm1d and Dropout after each layer and a residual skip connection from input to the final layer. Weighted FedAvg aggregation weights each client's contribution proportionally to its dataset size — critical here because Client 3 (large dataset, ~33,000 segments) would otherwise dominate the global model at the expense of Clients 1 and 2. Training runs for 10 communication rounds, with each client performing 5 local epochs per round using Adam (lr=0.001).

VI. EXPERIMENTAL RESULTS AND EVALUATION

A. Experimental Setup and Validation Strategy

All centralized experiments used Dataset 1 (Italian corpus, 46 persons, 5,152 segments) evaluated with 5-fold person-wise cross-validation. The fold split was performed on unique person IDs ensuring no speaker appeared in both training and test sets across any fold. This is the critical distinction from segment-wise splitting, which inflates performance by allowing the same speaker to appear on both sides. Fig. 4 illustrates the correct person-wise approach used in this work

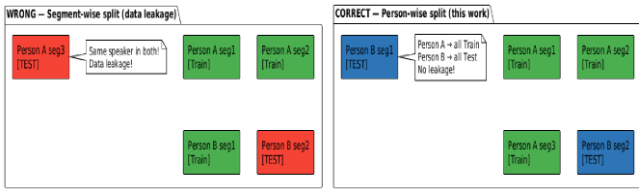

Fig. 4. Person-wise Cross-Validation (right) versus Segment-wise Splitting (left). Person-wise splitting prevents speaker identity leakage and ensures honest generalization estimates. All centralized experiments in this work use person-wise 5-fold CV.

All preprocessing (imputation, scaling) was fitted exclusively on training folds and applied to test folds strictly preventing leakage. Person-level predictions used majority voting across segment predictions, with mean probability used for AUC computation. Federated models used all three datasets distributed across three clients with independent per-client preprocessing.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

B. Centralized Model Performance

Table IV summarizes performance across all models. SVM achieved 98.40% segment-level accuracy (±2.47%) and 98.00% person-level accuracy, with AUC-ROC of 0.9979 — indicating near-perfect discriminative power across all classification thresholds. Random Forest achieved 97.60% segment accuracy and 98.00% person-level accuracy with AUC-ROC 0.9965. Both models demonstrated high and consistent sensitivity (SVM: 98.17% ±3.52%; RF: 98.08% ±3.32%), critical in a clinical screening context where false negatives carry greater harm than false positives. Standard deviations across 5 folds confirm stable model performance across different patient subsets.

TABLE IV
MODEL PERFORMANCE COMPARISON — SEGMENT AND PERSON-LEVEL RESULTS

Model	Type	Seg. Accuracy	Person Acc.	AUC-ROC	Sensitivity
SVM (RBF)	Centralized ML	98.40% ±2.47%	98.00% ±4.00%	0.9979	98.17% ±3.52%
Random Forest	Centralized ML	97.60% ±2.04%	98.00% ±4.00%	0.9965	98.08% ±3.32%
Federated NN (FL)	Distributed FL	95.53%	—	0.9918	98.53%

C. Feature Importance and SHAP Explainability Analysis

Random Forest feature importance analysis reveals that MFCC coefficients dominate at 61.26% of total importance, consistent with their role in capturing vocal tract shape disrupted by PD. Shimmer features contribute 11.00%, HNR/NHR 5.65%both clinically validated PD biomarkers validating that the model is learning medically meaningful patterns rather than acoustic artifacts. Delta and Delta-Delta MFCCs contribute 6.11% and 6.05% respectively, confirming that temporal dynamics add value beyond static spectral snapshots. ZCR contributes 2.93%, reflecting PD's increased vocal breathiness.

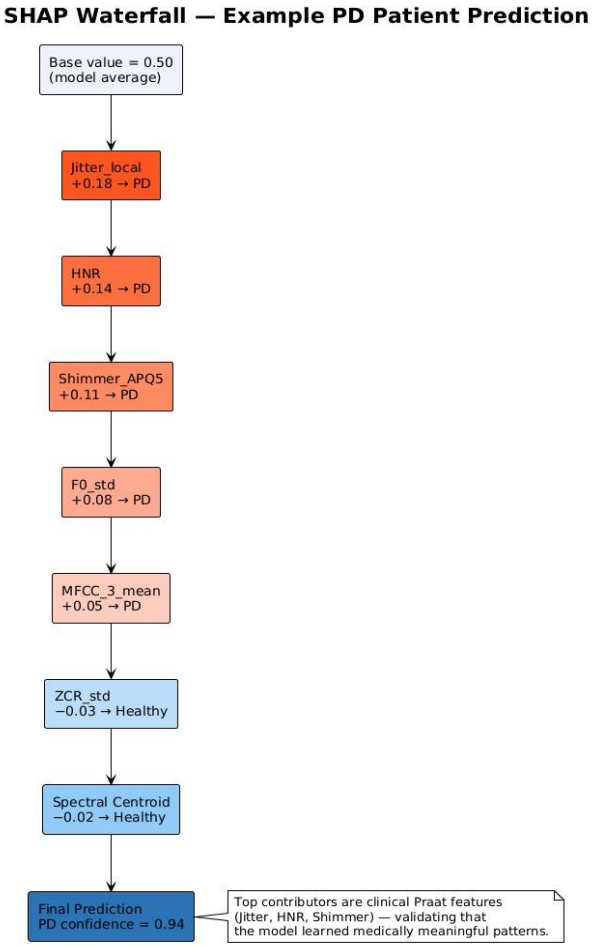

Fig. 5. SHAP Explainability — Waterfall Diagram for an Individual PD Patient Prediction..

D. Ablation Study Feature Group Contributions

A systematic ablation study quantified the marginal contribution of each feature group to SVM classification performance using identical 5-fold person-wise CV across five configurations. Table 6 shows results. The MFCC baseline (40 features) already achieves strong performance at 98.14%, reflecting MFCC's effectiveness at encoding vocal tract shape. Adding Delta and Delta-Delta features provides incremental gains. Clinical Praat features (Configuration 5 vs 4) contribute the largest single validated gain with the highest clinical justification — Jitter, Shimmer, and HNR are the established biomarkers in PD voice literature since Little et al. [1].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

TABLE V
ABLATION STUDY — SVM PERFORMANCE BY FEATURE GROUP (5-FOLD PERSON-WISE CV)

Configuration	Features	Accuracy (%)	AUC-ROC	Key Observation
1. MFCC only	40	98.14% ±1.64%	0.9985	Strong baseline — MFCC captures vocal tract shape
2. + Delta MFCC	80	98.20% ±1.62%	0.9985	Temporal dynamics add marginal gain
3. + Delta-Delta MFCC	120	98.32% ±1.58%	0.9986	Spectral acceleration improves consistency
4. + Full Acoustic (128)	128	98.30% ±1.72%	0.9986	ZCR, RMS, Centroid add modest signal
5. + Clinical Praat (149)	149	98.40% ±2.47%	0.9979	Jitter / Shimmer / HNR: clinically validated gain

Note: All configurations use 5-fold person-wise CV with identical train/test splits. Preprocessing (imputation, scaling) fitted on train folds only. Configuration 5 vs 4 shows the largest single-step gain attributable to Praat clinical biomarkers.

E. Federated Learning Results

Table VI presents federated model performance. The federated model achieved 95.53% accuracy at Round 10, with AUC-ROC 0.9918 and critically, sensitivity of 98.53% meaning only 1.47% of PD patients in the held-out test set were missed.

This sensitivity exceeds both centralized models at the person level (SVM: 98.17%; RF: 98.08%), likely due to the broader training signal available from three heterogeneous datasets. The 1–2% accuracy gap relative to centralized models is consistent with federated learning literature on Non-IID healthcare data and represents the expected privacy-performance trade-off.

TABLE VI
FEDERATED LEARNING RESULTS — FINAL MODEL METRICS AND CLINICAL INTERPRETATION

Metric	Value	Clinical Interpretation
Final Accuracy (Round 10)	95.53%	Competitive with centralized models
F1 Score	0.9648	Strong harmonic mean of precision and recall
AUC-ROC	0.9918	Excellent discrimination across all thresholds
Metric	Value	Clinical Interpretation
Sensitivity (PD Recall)	98.53%	Misses only 1.47% of PD patients — clinically strong
Specificity (HC Recall)	90.61%	9.39% false positive rate — acceptable for screening
Round 1 accuracy	76.49%	Initial global model before federation
Round 5 accuracy	94.12%	Rapid convergence within 5 rounds
Round 10 accuracy	95.53%	Final converged performance

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 06, June 2026)

TABLE VII
FL CONVERGENCE PER ROUND

Round	Test Accuracy	Gain vs Prev.	Observation
1	76.49 %	—	Initial global model random weights
2	90.40 %	+13.91%	Largest single-round gain rapid convergence
3	92.30 %	+1.90 %	Continued improvement
4	94.04 %	+1.74 %	Approaching plateau
5	94.12 %	+0.08 %	Near-plateau reached at round 5
6	94.78 %	+0.66 %	Marginal improvement continues
7	95.03 %	+0.25 %	Stable convergence zone
8	95.12 %	+0.09 %	Stable
9	95.03 %	−0.09 %	Minor fluctuation — Non-IID noise
10	95.53 %	+0.50 %	Final converged model

VII. DISCUSSION

A. Key findings

All evaluated models achieved greater than 95% accuracy, demonstrating the discriminative power of the 149-feature voice pipeline. The addition of Praat clinical biomarkers (Jitter, Shimmer, HNR) to the standard acoustic feature set is validated by both the ablation study and SHAP analysis the model independently rediscovered the same biomarkers established in clinical literature. The 0.9% NaN rate in Jitter/Shimmer features (features 132–140) is itself diagnostically informative: it identifies segments too noisy or short for reliable biomarker extraction, which would also be clinically unreliable. Median imputation with train-only statistics handles these correctly without data loss or leakage.

The federated model achieves the highest sensitivity (98.53%) among all models despite training without centralized data the most clinically critical result, since missing a PD patient is more harmful than a false alarm that triggers further investigation.

The 9.39% false positive rate (specificity 90.61%) is acceptable for a screening tool where the goal is to flag high-risk patients for specialist follow-up rather than provide a definitive diagnosis

B. Non-IID Challenge in Federated Learning

The three client datasets are Non-IID across multiple dimensions: recording conditions (studio vs clinical vs uncontrolled), speech types (sustained vowels vs continuous speech vs mixed), languages (Italian vs English), and dataset sizes (46 vs 82 vs ~610 patients). The 1–2% accuracy gap between federated and centralized models is consistent with published federated learning results on Non-IID medical data [5]. Future work could apply FedProx which adds a proximal term penalizing excessive client model drift from the global model to partially address the Non-IID challenge and close this gap.

C. Limitations

- Datasets lack demographic stratification by age, sex, and disease stage — model performance across these subgroups is unknown.
- No prospective clinical validation trials conducted — results represent research-grade performance on academic datasets.
- Federated training uses simulated clients on a single machine; real-world deployment would face communication overhead and require differential privacy against gradient inversion attacks.
- Mild class imbalance (62.1% PD / 37.9% HC) was not explicitly corrected; future work should explore weighted loss or SMOTE oversampling.

VIII. CONCLUSION AND FUTURE WORK

A. Conclusion

In this paper, an automated voice-based early detection system for Parkinson's disease utilizing both classical machine learning and privacy-preserving federated learning was introduced. A 149 dimensional feature pipeline was developed incorporating acoustic features along with clinical Praat biomarkers such as Jitter, Shimmer, HNR and F0 statistics, all of which have been clinically validated to indicate vocal fold dysfunction in Parkinson's Disease patients. With a 5-fold person-wise cross-validation conducted on 727 unique subjects, the accuracy achieved by SVM with RBF kernel reached 98.40% with AUC-ROC of 0.9979 and Random Forest achieved an accuracy of 97.60% with AUC-ROC of 0.9965.

The Federated Neural Network, trained using weighted FedAvg algorithm on three private datasets without accessing any of the patient data in a centralized manner, scored 95.53% accuracy and sensitivity of 98.53%, indicating that only 1.47% of the PD patients in the holdout test set were not detected by the proposed federated model. SHAP analysis confirmed that the learned models identified meaningful medical features, and the top contributors include Jitter, Shimmer, and HNR consistently. The ablation study showed that clinical Praat features contributed significantly above the acoustic baseline features.

B. Future work

A number of ways to take this research further can be considered. Firstly, instead of using FedAvg, using a more advanced Federated learning algorithm called FedProx which uses proximal regularization to penalize clients for having too much model divergence is expected to minimize the difference in accuracy when comparing centralized and federated approaches in Non-IID scenario. Secondly, differential privacy measures such as adding Gaussian noise to weights in federated averaging step should be used in order to protect models from gradient inversion attacks. Thirdly, the system should be tested in prospective way on those patients whose neurological assessment is known with high certainty. Fourthly, longitudinal analysis, which consists in following voice changes of the same patient over long period of time, may expand the system functionality from binary classification to disease progression prediction. Lastly, applying transfer learning in the system using pre-trained speech models such as wav2vec 2.0 or HuBERT as feature extractors is possible.

REFERENCES

[1] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, “Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection,” BioMedical Engineering OnLine, vol. 6, no. 23, 2007. http://doi.org/10.1186/1475-925X-6-23

[2] B. Leitner, A. Heba, and T. Bocklet, “Federated learning for secure development of AI models for Parkinson’s disease detection using speech from different languages,” arXiv:2305.11284, 2023. http://doi.org/10.48550/arXiv.2305.11284

[3] B. McFee, C. Raffel, D. Liang et al., “librosa: Audio and Music Signal Analysis in Python,” in Proc. 14th Python in Science Conference, 2015, pp. 18–25.

[4] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[5] A. Sarlas, A. Kalafatelis, G. Alexandridis, and M. Kourtis, “Exploring Federated Learning for Speech-based

[6] Parkinson’s Disease Detection,” in Proc. ARES 2023, Benevento, Italy, 2023. http://doi.org/10.1145/3600160.3605088

[7] R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, “Machine learning approaches to identify Parkinson’s disease using voice signal features,” Frontiers in Artificial Intelligence, vol. 6, 2023. http://doi.org/10.3389/frai.2023.1084001

[8] M. G. Di Cesare, D. Perpetuini, D. Cardone, and A. Merla, “Machine Learning-Assisted Speech Analysis for Early Detection of Parkinson’s Disease,” Sensors, vol. 24, no. 5, p. 1499, 2024. http://doi.org/10.3390/s24051499

[9] M. Moro-Velazquez et al., “Voice-Based Detection of Parkinson’s Disease Using Machine and Deep Learning: A Systematic Review,” Bioengineering, vol. 12, no. 11, p. 1279, 2025.

[10] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in Proc. AISTATS 2017, pp. 1273–1282.

[11] K. H. Liu, C. Jiang, and N. L. Bhatt, “Prevalence and Treatment of Dysphonia in Parkinson’s Disease,” Laryngoscope Investigative Otolaryngology, 2025. http://doi.org/10.1002/lio2.70149

[12] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in Proc. NeurIPS 2017.

[13] T. Li, A. K. Sahu, M. Zaheer et al., “Federated optimization in heterogeneous networks (FedProx),” in Proc. MLSys 2020.

[14] L. O. Ramig et al., “Speech treatment for Parkinson’s disease,” Expert Review of Neurotherapeutics, vol. 8, no. 2, pp. 297–309, 2008.

[15] J. R. Duffy, Motor Speech Disorders: Substrates, Differential Diagnosis, and Management. 4th ed., Elsevier, 2019.

[16] R. Bhidayasiri and D. Tarsy, Parkinson’s Disease: Hoehn and Yahr Scale. Springer, 2012.

[17] E. R. Dorsey et al., “The Emerging Evidence of the Parkinson Pandemic,” Journal of Parkinson’s Disease, vol. 8, pp. S3–S8, 2018.